Nouveau Groupe de l'Institut Fredrik Bull

« Enjeux épistémologiques et éthiques de l'apprentissage machine profond »

Proposé par Christophe Denis, Maître de Conférences HDR à La Sorbonne
18 Avril 2023

Résumé des motivations

Le retour tonitruant des réseaux de neurones s'est produit dans le sublime décor florentin en 2012 lors d'une conférence internationale renommée sur la vision par ordinateur. D'autres disciplines scientifiques computationnelles, comme la mécanique des fluides, la géophysique et la climatologie, ont également commencé à utiliser des méthodes d'apprentissage en profondeur pour prédire des phénomènes difficiles à résoudre avec une approche déductive hypothétique classique.
Impressionné par les résultats obtenus par les résultats du deep learning (apprentissage profond) , un étudiant américain en master de l'université du Maryland avait mis en place un ambitieux projet de deep learning. L'élève et son professeur ont été bluffés par les très bons résultats obtenus par le modèle... jusqu'à ce qu'un husky dans la neige soit classé comme un loup par le réseau de neurones profonds. Après une analyse plus approfondie, l'explication des très bons résultats de prédiction a été décevante : le réseau de neurones n'a pas "appris" à distinguer un loup d'un husky, mais seulement à détecter les décors enneigés. Le modèle d'apprentissage automatique a-t-il triché ?

Alors, comment établir la confiance entre les utilisateurs et l'IA ? Pour garantir la confiance, de nombreux comités d'éthique de l'IA recommandent d'intégrer des explications sur les résultats prédictifs de l'apprentissage automatique à fournir aux utilisateurs. Par exemple, en France, la loi de bioéthique récemment votée par l'Assemblée nationale française impose aux concepteurs d'un dispositif médical basé sur le machine learning d'en expliquer le fonctionnement.

La multiplication des comités d'éthique autour de l'IA cache une crise des fondements de l'informatique théorique comme cela a été le cas pour la physique et les mathématiques au début du vingtième siècle. La demande d'explication est légitime aux niveaux des sciences humaines et sociales mais ne doit pas masquer le besoin de revisiter la théorie de l'information (Fisher, Shannon) et la philosophie de l'information (ex. Floridi) à la lumière de l'avènement d'un monde post cybernétique pour définir un cadre de contrôle et de validation. Il s'agit de produire des informations permettant aux utilisateurs de produire leurs propres raisonnements plutôt que de recevoir passivement une explication.

Ce cadre de validation ne pas éclipser les nouveaux enjeux épistémologiques et éthiques liées à la mise au point de techniques d'apprentissage machine plus sophistiquées sur lesquelles travaillent notamment les équipes de Le Cun au sein du laboratoire d'Intelligence Artificielle de Facebook. En effet, ces méthodes d'apprentissage machine auto-supervisées, basées sur des concepts de physique statistique, leur permet d'acquérir une capacité de généralisation grâce à l'apprentissage d'une certaine représentation unifiée du monde. Cela permet par exemple pratiquement d'annoter automatiquement de nouvelles données en se basant sur un critère énergétique que l'on peut rapprocher de ceux utilisés en théorie de l'information.

Il s'agit de proposer une grille de lecture épistémologiques pour se prémunir d'effets d'annonces commerciales, de mesurer ses implications, positives ou négatives, sur l'intégrité de la connaissance et la découverte scientifiques et enfin de proposer une méthodologie de contrôle cybernétique permettant de contrôler l'opacité des ces boites noires que nous démontrerons

nécessaire d'un point de vue éthique.

Idée d'organisation du Groupe

La durée du groupe de travail est initialement de deux ans.

1. La première année sera consacrée à la définition d'un programme de recherche pour redéfinir la théorie de l'information et de la simulation dans une approche cybernétique en prenant des modèles d'apprentissage profond machine supervisé

2. La seconde année du GT sera proposer d'appliquer cette méthodologie et de raffiner cette méthodologie sur des méthodes d'apprentissage machine plus sophistiquées comme les modèles autosupervisés.

Livrable du GT :
Le GT produira un rapport intermédiaire au bout de la première année et un rapport consolidé à la fin du GT. Ce rapport sera publié dans un ouvrage collectif.

Constitution du comité de pilotage
Le comité de pilotage du groupe de travail constitué d'un(e) animateur(e), d'un(e) co-animateure et de deux autres personnes. Ce comité sera chargé de définir le calendrier ses séances de travail, puis de proposer une liste des questions à soumettre au prochain conférencier. Ce comite sera également en charge de produire les deux rapports et de gérer la publication de l'ouvrage collectif. L'équipe d'animation sera en charge de piloter ce comité en dehors des séances de travail.

Une plateforme de type Slackware sera mis en place pour permettre au comité de pilotage de communiquer et de s'organiser entre les séances.